Comparative Analysis of Predictive Modeling Techniques for Archaeological Site Prediction in Upper Galilee and Fuxin Regions

Shyam Sankar
ECEN 649
Texas A&M University
College Station, TX 77843
shyam.sankar@tamu.edu

Abstract

This project aims to compare Archaeological Predictive Models (APM) by implementing and comparing two pattern recognition techniques: logistic regression and decision trees. Using datasets from two culturally and geographically distinct regions - Upper Galilee in northern Israel and the Fuxin area in northeast China - this study will investigate the potential of each model in accurately predicting the likelihood of archaeological sites across varied landscapes. The models' performances will be assessed to determine the most effective methodology for spatial prediction in archaeological contexts. The focus of evaluating the models will be on its ability to detect archaeological sites effectively.

1 Introduction

An Archaeological Predictive Model (APM) is a machine learning model that uses data collected on geographical and environmental factors to predict the likelihood of finding an archaeological site in a particular area. APMs are commonly used to understand the archaeological significance of a location before an activity like a construction project is undertaken, avoiding the potential destruction of archaeological evidence.[1]

A simple way of looking at the problem statement is to pose the question "How would we identify this given region as having had human presence in the past?". Intuition would tell us that human beings would have a tendency to settle in regions that would support their life. And often in historic times, this meant that surrounding nature should have enough resources to sustain the settling population.

This leads us to ask the question "Does this region have enough natural resources that can sustain a settlement of people over time?". Resources in this context would mainly refer to food and water. We could say that the presence of a fresh water source like rivers and lakes was crucial. A water source often served as a food source as fishing was a common practice. However, most settlements would look for multiple food sources, a site that is close to abundant wildlife suitable for hunting would be key for a steady supply of meat. Areas with rich plant life, that could provide fruits, berries and nuts would have been ideal.

The presence of archaeological sites as a result of human settlement is also subject to other factors like weather patterns, topography and geological features. Settlements would avoid extreme climates and prefer moderate climates. Preferences for elevated regions is natural to allow for defense against predators and better visibility. The presence of stones to make stone tools for hunting resources for making shelters would have been highly sought out.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

2 Background

In 1970, Vita-Finzi, C. et al. tried to establish a relationship between human settlements, growth of economy and environmental factors, citing the availability of resources and their exploitation played a huge role in development of human presence in a prehistoric region.

This was followed by Plog, F. et al. (1971) where mathematical models where used to predict the presence of archaeological sites. This provided archaeologists with a one of a kind tool that can be used to selectively survey regions after predicting the likelihood of human presence in the region.

The 1980s saw the development of quantitative models, where environmental data collected from archaeological sites where used to develop assessments that could be used in areas that are yet to be surveyed. This laid the foundation for application of models using pattern recognition techniques to survey regions for detecting the presence of archeological sites.

3 Data Sets

3.1 Upper Galilee

The first of the two data sets being used was collected from the Upper Galilee region of Northern Israel. The region has an elevation of 1200 meters and forms a part of Israel's central mountain range. The region has an average annual precipitation of 960 mm on the Meron ridge and 550-600 mm in other areas. The region is suitable for growing winter crops like wheat and barley. The region has limited water resources, mostly comprising of small springs. However, the steady precipitation rate would mean cisterns where used to collect rain water.

The data used in this project is from the work of Frank et al. (2001), who surveyed and collected data from 54 habitation sites, previously identified as Bronze and Iron age settlement sites. A total area of 378 square kilometers was studied.

3.2 Fuxin

The Fuxin area in the Liaoning Province of North China is a moderately hilly area with elevations reaching up to 390 meters is the ridges and 140 to 20 meters above the sea level in the valleys. The area has an average annual precipitation rate of 400 mm. The valleyed regions were suitable for agriculture with areas often used as fruit orchards.

The data used in this project is from the work of Shelach-Lavi et al. (2016), with data coming from survey conducted of 104 square kilometers. The habitation sites were previously identified as regions with habitation sites dating back to the Bronze and Neolithic age.

3.3 Environmental Variables

The environmental features available for the Upper Galilee region are: distance from Chalk rock formations, aspect, land curvature, distance from springs, distance from valley floors, distance from agricultural land, distance from grazing land.

The environmental features available for the Fuxin region are: distance from agricultural lands, distance from modern villages, distance from forests, slope, aspect, land curvature, elevation and main river distance.

4 Methods

4.1 Logistic Regression

The most commonly used pattern recognition technique for developing an Archaeological Predictive Model is Logistic Regression. With the logistic regression model, we predict the probability of the presence of an archaeological site in a given coordinate. The presence or absence of an archaeological site serves as the dependent variable while the environmental factors serves as independent variable.

Logistic regression attempts to find the best fitting line that establishes a relationship between the independent variables and dependent variables. This relationship is used to predict the probability of the presence of an even like the presence of archaeological sites from the features.

For example, if we are using only elevation and slope as features for the model, the model would look like this

Logit(P) = $b_0 + b_1(elevation) + b_2(slope)$

The coefficients are estimated using maximum likelihood estimation.

Separate models were trained and tested for both the Galilee and Fuxin regions.

4.2 Decision Trees

The project aims to build upon the work of Wachtel et al. (2018) by applying decision trees to classify regions.

Decision trees are a pattern recognition technique that recursively slits the data by selecting the best features at each node. To select which features to split at each node, a decision tree model uses information gain or gini index. Once a tree is trained and built, it can be used to make predictions.

We have decided to use decision trees because of the high interpret-ability of the model. We can use decision trees to come up with conclusions like "if distance from water sources is less than 100 meters and elevation is greater than 500 meters, the probability of an archaeological site is high".

4.3 Evaluation Metrics

Evaluation of archaeological prediction models proposed will be made using metrics that would assess how well the model is capable of identifying archaeological sites while trying to minimize false positives and false negatives. We will use precision, recall, F-1 score and the ROC-AUC curve to evaluate the model, with a particular focus on the recall values for known sites.

4.3.1 Precision

Prediction of non-archaeological sites as archaeological sites can lead to a waste of valuable resources as excavation could probably be called for the region. Precision is a measure that indicates the proportion of true positives out of all the predicted positives. In our context, it is proportion of correctly predicted archaeological sites out of all sites that were predicted to be archaeologically significant.

$$Precision = \frac{TruePositives(TP)}{TruePositives(TP) + FalsePositives(FP)}$$

4.3.2 Recall

Recall is a measure that indicates the proportion of true positives that were predicted by the model. Recall is a measure of how well the model is capable of predicting archaeological sites. In the context of this project, recall could be called the most important measure of the archaeological predictive model, if it is achieved with considerable precision. APMs missing on any archaeological sites can lead to missed opportunity for research and discovery, hence we will prioritize a high recall value.

$$Recall = \frac{TruePositives(TP)}{TruePositives(TP) + FalseNegatives(FN)}$$

4.3.3 F1 Score

F1 score is the harmonic mean of recision and recall. It is a measure of balance between precision and recall.

$$F1Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

4.3.4 Receiver Operating Characteristic - Area Under the Curve

The ROC-AUC is a common evaluation metric that shows how well a binary classification model distinguishes between the two classes, across different thresholds. The ROC curve is a plot between the True Positive Rate and the False Positive Rate. The area under this curve can be used as a performance metric. A model that has an AUC of 0.5 indicates that the model is randomly guessing between the classes, while an AUC value of 1 indicates that the model perfectly distinguishes between the two classes.

5 Experimental Results

5.1 Key Features

From running the model on the data collected from Galilee region, we observe from the coefficients of the logistic regression model that, environmental features like distance from water springs, slope and land curvature are highly associated with the presence of archaeological sites.

Table 1: Feature Names and Their Absolute Coefficients for Upper Galilee

Feature Name	Absolute Coefficient
Distance from springs	1.033479
Slope	0.994676
Land Curvature	0.790803
Distance from chalk rocks	0.478908
Distance from agricultural land	0.359164
Distance from valley floor	0.276669
Distance from grazing land	0.050621
aspect	0.006377

From running the model on the data collected from Fuxin region, we observe from the coefficients of the logistic regression model that, environmental features like distance from river, distance to crops and elevation are highly associated with the presence of archaeological sites.

Table 2: Feature Names and Their Absolute Coefficients for Fuxin

Feature Name	Absolute Coefficient
Distance to river	0.950976
Distance to crops	0.566248
Elevation	0.563241
Distance to forests	0.359826
Distance to Modern Buildings	0.139460
Aspect	0.058205
Land curvature	0.030745
Slope	0.029668

5.2 Evaluation

A good Archaeological Predictive Model should be able to detect as many archaeological sites as possible. Hence, the focus of the model is to successfully detect as many of the true archaeological sites as possible. Hence, a focus on the recall values of sites for model evaluation.

5.2.1 Upper-Galilee

Table 3: Precision and Recall for Decision Tree and Logistic Regression Models

Model	Class	Precision	Recall	F1 Score
Logistic Regression	Non-sites	0.79	0.54	0.64
	Sites	0.42	0.70	0.53
Decision Tree	Non-sites	0.86	0.75	0.80
	Sites	0.59	0.74	0.65

From the above table, we can gather that logistic regression successfully detects 70% of the archaeological sites, however, a low precision value of 0.42 indicates that a large number of false positives, sites that are marked as potential archaeological sites when they are not. We further look into decision trees to filter out some of the false positives. The AUC for the logistic regression model is 0.68.

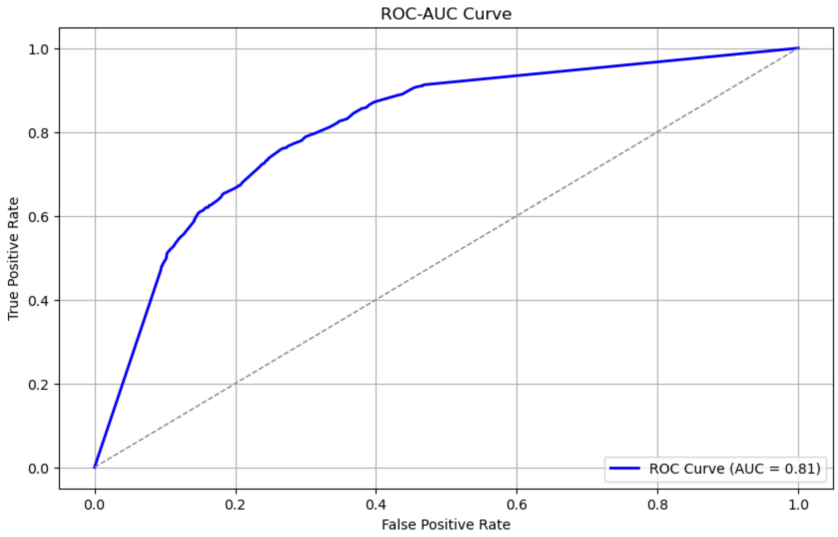

Figure 1: ROC-AUC Curve for Decision tree classifier - Upper Galilee

The parameters for the decision tree are set with the criterion as 'entropy,' a maximum depth of 20, a minimum of 5 samples per leaf, and a minimum of 2 samples required for a split.

The decision tree classifier classifies 74% of the archaeological sites accurately. While this measure is lower than the logistic regression classifier, a precision of 0.59 is an improvement upon the logistic regression model. The AUC for the decision tree classifier is 0.81, a significant improvement in comparison to the logistic regression model.

5.2.2 Fuxin

Table 4: Precision and Recall for Decision Tree and Logistic Regression Models

Model	Class	Precision	Recall	F1 Score
Logistic Regression	Non-sites	0.94	0.43	0.59
	Sites	0.38	0.93	0.54
Decision Tree	Non-sites	0.93	0.82	0.87
	Sites	0.64	0.83	0.72

From the above table, we can gather that logistic regression successfully detects 93% of the archaeological sites, however, a low precision value of 0.38 indicates that a large number of false positives, sites that are marked as potential archaeological sites when they are not. We further look into decision trees to filter out some of the false positives. The AUC for the logistic regression model is 0.71.

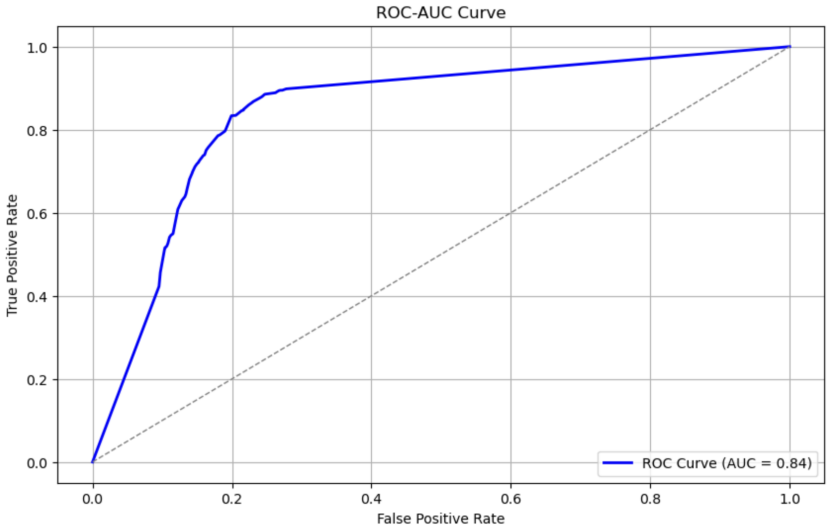

Figure 2: ROC-AUC Curve for Decision tree classifier - Fuxin

The decision tree parameters are configured with the criterion set to 'gini,' a maximum depth of 15, a minimum of 1 sample per leaf, and a minimum of 2 samples required for a split.

The decision tree classifier classifies 83% of the archaeological sites accurately. While this measure is lower than the logistic regression classifier, a precision of 0.64 is an improvement upon the logistic regression model. The AUC for the decision tree classifier is 0.84, a significant improvement in comparison to the logistic regression model.

6 Conclusion

Environmental features like distance from water springs, slope, and land curvature are highly associated with the presence of archaeological sites in the Galilee region of Northern Israel because they influenced ancient settlement patterns and resource use, especially around the Bronze and Iron age. Proximity to water springs provided essential drinking water and agricultural opportunities, while gentle slopes facilitated farming and strategic settlement for defense purposes. Similarly, land curvature affected natural drainage, agricultural potential, and shelter from environmental challenges, making these areas favorable for habitation. .

In the Fuxin region of Liaoning province in Northern China, environmental features such as distance from rivers, distance to crops, and elevation are highly associated with the presence of archaeological sites because they directly impacted ancient human settlement and resource management. Proximity to rivers provided essential water sources for drinking, agriculture, and transportation, while proximity to crops indicates the agricultural dependence of ancient societies. Elevation influenced settlement choice by offering strategic advantages for defense, as well as suitability for farming and minimizing the risk of flooding.

From the above experimental results, we can conclude that decision trees are an effective pattern recognition technique for building archaeological predictive models. We establish that decision trees provide an improvement in most evaluation metrics when compared to the commonly used logistic regression models.

A common challenge faced by state-of-the-art archaeological predictive models are the poor generalization of models to other areas. This is partially due to how different each region is, but also due

to the lack of resources, limiting archaeological studies from conducting large scale excavation and surveys. While our model achieves considerably good accuracy, the model could be improved further if there was availability of more data for different regions surrounding these settlements.

GitHub Link: https://github.com/shyamsankar11102000/ArchaeologicalPredictiveModels

References

[1] Whitley, Thomas. 2020. "An Introduction to Archaeological Predictive Modeling" - SAA Webinar. 10.13140/RG.2.2.19242.16326.

[2] Vita-Finzi, C. et al. 1970. “Prehistoric Economy in the Mount Carmel Area of Palestine: Site Catchment Analysis.” Proceedings of the Prehistoric Society 36: 1–37. doi: 10.1017/S0079497X00013074..

[3] Plog, F., & Hill, J. M. 1971. "Explaining Variability in the Distribution of Sites. In G. J. Gumerman (ed.), The Distribution of Prehistoric Population Aggregates". Prescott: Prescott College Press, pp. 7–37.

[4]Frankel, R., et al. 2001 "Settlement Dynamics and Regional Diversity in Ancient Upper Galilee (Israel Antiquities Authority Reports 14).".

[5] Shelach-Lavi, Gideon, et al. 2016 "Human adaptation and socioeconomic change in northeast China: Results of the Fuxin Regional Survey." Journal of Field Archaeology 41.4 : 467-485.

[6] Wachtel, Ido & Zidon, Royi & Garti, Shimon & Shelach-Lavi, Gideon. 2018. "Predictive modeling for archaeological site locations: Comparing logistic regression and maximal entropy in north Israel and north-east China". Journal of Archaeological Science. 92. 22-36. 10.1016/j.jas.2018.02.001.